Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 78 (2016) 120 – 124
Procedia Computer Science

International Conference on Information Security & Privacy (ICISP2015), 11-12 December 2015, Nagpur, INDIA

Privacy Preserving Unstructured Big Data Analytics: Issues and Challenges

Brijesh B. Mehta*, Udai Pratap Rao

Computer Engineering Department, S. V. National Institute of Technology, Surat-395007, India

Abstract

Big data analytics has created opportunities for researchers to process huge amount of data but created a big threat to privacy of individual. Data processed by big data analytics platforms may have personal information which need to be taken care of when deriving some useful results for research. Existing privacy preserving techniques like, anonymization requires having dataset divided in the set of attributes like, sensitive attributes, quasi identifiers, and non-sensitive attributes. With the structured data it may possible to have such a distribution but in unstructured data it is very difficult to identify sensitive attribute and quasi identifiers.

Peer-review under responsibility of organizing committee of the ICISP2015

Keywords: Big Data; Privacy; Unstructured Data

1. Introduction

Big data can be defined as, “The data sets so large or complex that are difficult to process using traditional data processing applications”[1]. Size of big data may be in zeta bytes (increasing proportionally with time). Big data has characteristics of 3Vs; Volume (large amount of data), Velocity (speed of data generation and processing), and Variety (structured, unstructured, or semi-structured data).

* Corresponding author.
E-mail address: brijeshbmehta@acm.org

Peer-review under responsibility of organizing committee of the ICISP2015
doi:10.1016/j.procs.2016.02.020

Big data analytics is very helpful in various fields like, medical science, national security, semantic web, social media, etc. On the other hand, it creates a privacy threat to an individual as it has capacity to store and process large amount of data very quickly and accurately, due to advancement in technologies like, NoSQL data models[2, 3, 4], Hadoop[5], Map-reduce[6], etc. Therefore, instead of seeing the picture as big data analytics vs. privacy, we need to have individual privacy preservation with almost all advantages of big data analytics. A privacy preserving technique is required which maintains a trade-off between privacy and utility of individual's data.

To understand the importance of privacy in big data analytics, privacy issues with big data analytics have been discussed in next section. Why existing privacy preserving models can't be used for big data analytics? Answers for this question have been given in the challenges section.

2. Privacy Issues With Big Data Analytics

In this section, discussion about the privacy issues in various areas are given.

2.1. Privacy Issues in Big Mobile Data

Everything is available on mobile nowadays. People are sharing lot of information on mobile phones. Often, mobile sends data to the service provider without user's knowledge. Identifying the person using his mobile data and the details provided by the service provider is very easy. Therefore, privacy in mobile data is very important.

Lee Garber[7] has mention that, "Bit Defender, a Romanian security vendor, has found many Android applications, which can access and even send information like, user's location, contact lists with phone numbers and email addresses, as well as photos without their consent. Bit Defender also analyzed 836,021 Android applications on Google Play Store and found that about 33% of apps could reveal location-related data. About 5% located and opened photos on user's phones, and approximately 3% reveal their email data."

Mirco Musolesi[8] has mention that "In June 2013, Facebook had, on average, 819 million monthly active mobile users." This is about Facebook only but there are many other mobile apps too (i.e., Twitter, WhatsApp, etc.) from which huge amount of data is being generated.

Text message analysis is an example of unstructured big data analytics in mobile. Mobile application like WhatsApp is using text message analysis in their mobile number verification method. In such method, a verification number sent to the registered mobile number and if it is same in which app is installed, app will automatically write the number in verification box as soon as it arrives via SMS.

2.2. Privacy Issues in Health Care Data

Patil et al.[9] have discussed some of the security and privacy issues. Big data analytics and genome research having real time access to patient record helps doctors to take decisions. Electronic Health Records (EHR)[10] helped a lot to digitize the health care system and EHR incentive program[11] motivates hospitals to create an accurate and complete EHR. On the other hand EHR having personal information of patient may lead to privacy breach. Therefore, privacy preserving analysis of data is required and data need to be anonymized or encrypted before data analysis.

Pathology report analysis is an example of unstructured data analysis in health care. Pathology reports are mostly in text paragraph form. Therefore, it is very difficult to identify sensitive attributes in such reports.

2.3. Privacy Issues in Social Media Data

Social media is one of the biggest revolutions in past decade. Lot of information is being shared by people on social media. Sometimes, people close to you shares some information about you, which you don't want disclose on social media. This may lead to privacy violation of an individual. For e.g.; you have taken a sick leave from office to watch a football game and one of your friend checked-in[12], you on facebook so people comes to know that you were not sick but having fun. Though, privacy settings are there in facebook to approve tag so if someone tagged you, it

need to be approved before it is posted on your wall but it is going to appear on your friend's wall as soon as he is posts with link to your profile.

Facebook status analysis or twitter's tweet analysis is the example of unstructured big data analytics in social media.

2.4. Privacy Issues in Web Usage Data

Sedayao et al.[13] have mentioned a real time scenario of web usage data mining. Intel want to make its internal website dynamic (appearance of the website changes as per the access pattern of users, viz. links visited by most number of users should be on the first page to save click time and improve productivity) based on web usage data of all the users of the website. With browser information and IP address from web usage data any user can be identified and whatever activities he is performing on line may be detected. Therefore, user privacy is violated by such system. Sedayao et al. have suggested a model in which symmetric key encryption is used to anonymize the sensitive data identified based on predefined tags like, IP address and user ID from semi-structured web usage data.

3. Research Challenges

In this section, research challenges with existing privacy preserving techniques for unstructured big data have been discussed.

3.1. Privacy Preserving Data Publishing Techniques

Fung et al.[14] have made a survey of some of the privacy preserving data publishing techniques (PPDP).

In PPDP, attributes are classified as[14, 15]:

- Personal Information Identifier (PII): Attributes that can directly and uniquely identify an individual, such as name, ID number and mobile number.
- Quasi-identifier (QID): Attributes that can be linked with external data to re-identify individual records, such as gender, age and zip code.
- Sensitive Attribute (SA): Attributes that an individual doesn't want to disclose, such as disease and salary.
- Non-sensitive Attribute: All attributes other than ID, QID and SA.

To preserve privacy, different anonymization operation[14] may apply on above mention attributes like:

- Generalization: replace the original value with generalize value of data from the same class like; replace male/female with gender.
- Suppression: replace original value of data with some special character like, *. For e.g.; value of pin code 395007 suppressed to 395***.
- Anatomization: putting QID and SA in different tables to break the relation between them.
- Permutation: creating groups based on QID and then shuffle the values of SA in each group to remove relation between QID and SA.
- Perturbation: replacing original value of SA data with some fake value.

PPDP techniques can not directly applied to big data because it is very difficult to differentiate sensitive and non-sensitive attributes in unstructured data. Therefore, it is a challenge for researchers to find out some technique which can be applied to both structured and unstructured data.

3.2. Unstructured Data De-identification Techniques

Gardner et al.[16] have shown in one of their paper that, with the help of conditional random field (CRF)[17], it is possible to identify sensitive attributes from the given unstructured data. There are mainly three phase of de-identifying unstructured data[16]:

- Identifying and sensitive information extraction, extracting identifying and sensitive information from unstructured data with the help of classification methods like, Bayesian classifier, conditional random field, etc.

- Data linking, Identifying and removing data link between sensitive and non-sensitive attributes (quasi identifiers).
- Anonymization, De-identifying data using techniques like, suppression, generalization, etc.

But the limitation with this method is that, it is not feasible to use probability based Bayesian classifier on large amount of data.

Thavavel et al.[18] have proposed a method of converting unstructured data to structured data using node representation. As shown in fig-1, authors have converted unstructured data to XML (semi structured data) and then map that XML file to node representation, finally get structured data as output. But here again the issue is with the size of data, converting large amount of unstructured data to structured is not feasible.

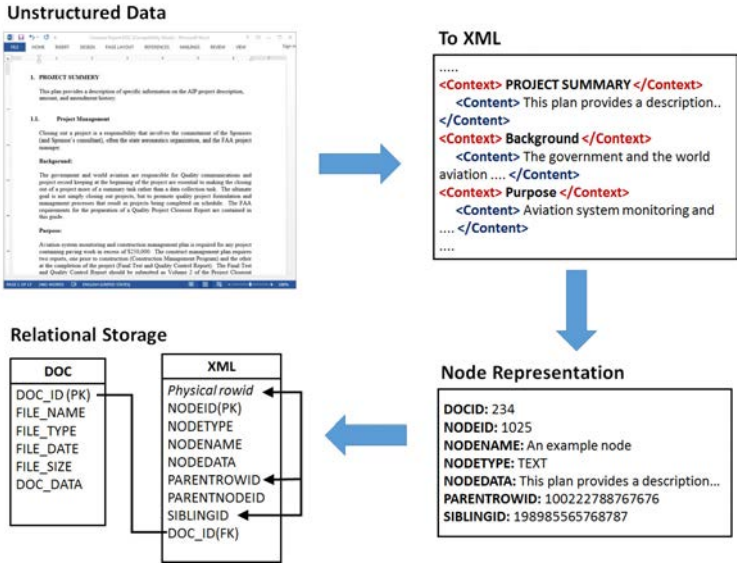

Fig 1. Converting unstructured data to structured data[18]

4. Conclusion

In this paper, discussion started from basics of big data and big data privacy. Privacy issues with big data analytics in different areas like, mobile data, health care, social media, and web usage data mining have been briefly explained. Research challenges with existing privacy preserving techniques for unstructured big data have been discussed. From above discussion we have found that existing privacy preserving techniques are able to address any one 'V' of 3Vs of big data but they fail for more than one 'V'. Therefore, a novel approach is required which can address volume as well as variety of data. In future, we are going to propose such a novel approach, which can extract a sensitive attributes from unstructured data using some of the machine learning techniques to preserve privacy of an individual.

References

1. "Big data definition," [Online] Available: http://en.wikipedia.org/wiki/Big data, [Accessed: 04-Nov-2014].
2. "Nosql database list," [Online] Available: http://nosql-database.org/, [Accessed: 04-Feb-2015].
3. K. Grolinger, W. A. Higashino, A. Tiwari, and M. A. Capretz, "Data management in cloud environments: Nosql and newsql data stores," Journal of Cloud Computing: Advances, Systems and Applications, vol. 2, no. 1, p. 22, 2013.
4. P. J. Sadalage and M. Fowler, NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence, 1st ed. Addison-Wesley Professional, 2012.
5. "Hadoop," [Online] Available: http://hadoop.apache.org/, [Accessed: 04-Nov-2014].

6. J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," Communications of the ACM, vol. 51, no. 1, pp. 107–113, Jan. 2008.
7. L. Garber, "Security, privacy, policy, and dependability roundup," IEEE Security and Privacy, vol. 11, no. 2, pp. 6–7, Mar. 2013.
8. M. Musolesi, "Big mobile data mining: Good or evil?" IEEE Internet Computing, vol. 18, no. 1, pp. 78–81, Jan. 2014.
9. K. Patil.H. and R. Seshadri, "Big data security and privacy issues in healthcare," in 2014 IEEE International Congress on Big Data (BigData Congress), June 2014, pp. 762–765.
10. "Ehr," [Online] Available: http://www.himss.org/library/ehr/, [Accessed: 25-Feb-2015].
11. "Ehr incentive programs," [Online] Available: http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Basics.html, [Accessed: 10-Feb-2015].
12. "Facebook checked in help," [Online] Available: https://www.facebook.com/help/461075590584469/, [Accessed: 10-Feb-2015].
13. J. Sedayao, R. Bhardwaj, and N. Gorade, "Making big data, privacy, and anonymization work together in the enterprise: Experiences and issues," in 2014 IEEE International Congress on Big Data (BigData Congress), June 2014, pp.601–607.
14. B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy preserving data publishing: A survey of recent developments," ACM Computing Surveys, vol. 42, no. 4, pp. 14:1–14:53, Jun. 2010.
15. Q. Wang, Z. Xu, and S. Qu, "An enhanced kanonymity model against homogeneity attack," Journal of Software, vol. 6, no. 10, 2011, [Online] Available: http://ojs.academypublisher.com/index.php/jsw/article/view/ jsw061019451952, [Accessed: 25-Feb-2014].
16. J. Gardner and L. Xiong, "An integrated framework for deidentifying unstructured medical data," Data and Knowledge Engineering, vol. 68, no. 12, pp. 1441–1451, Dec. 2009.
17. J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 282–289.
18. V. Thavavel and S. Sivakumar, "A generalized framework of privacy preservation in distributed data mining for unstructured data environment," International Journal of Computer Science Issues, vol. 9, no. 1, pp. 434–441, 2012, [Online] Available: http://www.ijcsi.org/papers/IJCSI-9-1-2-434-441.pdf, [Accessed: 24-Feb-2015].